

# Computing longest single-arm-gapped palindromes in a string

Shintaro Narisada<sup>1</sup>, Diptarama<sup>1</sup>, Kazuyuki Narisawa<sup>1</sup>, Shunsuke Inenaga<sup>2</sup>, and  
Ayumi Shinohara<sup>1</sup>

<sup>1</sup>Graduate School of Information Sciences, Tohoku University, Japan  
{shintaro\_narisada, diptarama}@shino.ecei.tohoku.ac.jp  
{narisawa, ayumi}@ecei.tohoku.ac.jp

<sup>2</sup>Department of Informatics, Kyushu University, Japan  
inenaga@inf.kyushu-u.ac.jp

## Abstract

In this paper, we introduce new types of approximate palindromes called *single-arm-gapped palindromes* (shortly *SAGPs*). A SAGP contains a gap in either its left or right arm, which is in the form of either  $wgucu^Rw^R$  or  $wucu^Rgw^R$ , where  $w$  and  $u$  are non-empty strings,  $w^R$  and  $u^R$  are respectively the reversed strings of  $w$  and  $u$ ,  $g$  is a gap, and  $c$  is either a single character or the empty string. We classify SAGPs into two groups: those which have  $ucu^R$  as a maximal palindrome (type-1), and the others (type-2). We propose several algorithms to compute type-1 SAGPs with longest arms occurring in a given string, based on suffix arrays. Then, we propose a linear-time algorithm to compute all type-1 SAGPs with longest arms, based on suffix trees. Also, we show how to compute type-2 SAGPs with longest arms in linear time. We also perform some preliminary experiments to show practical performances of the proposed methods.

## 1 Introduction

A palindrome is a string that reads the same forward and backward. Discovering palindromic structures in strings is a classical, yet important task in combinatorics on words and string algorithmics (e.g., see [4, 8, 13, 1]). A natural extension to palindromes is to allow for a *gap* between the left and right arms of palindromes. Namely, a string  $x$  is called a gapped palindrome if  $x = wgw^R$  for some strings  $w, g$  with  $|w| \geq 1$  and  $|g| \geq 0$ . Finding gapped palindromes has applications in bioinformatics, such as finding secondary structures of RNA sequences called *hairpins* [9]. If we further allow for another gap *inside either arm*, then such a palindrome can be written as  $wg_2ug_1u^Rw^R$  or  $wug_1u^Rg_2w^R$  for some strings  $w, g_1, g_2, u$  with  $|u| \geq 1$ ,  $|g_1| \geq 0$ ,  $|g_2| \geq 0$ , and  $|w| \geq 1$ . These types of palindromes characterize *hairpins with bulges* in RNA sequences, known to occur frequently in the secondary structures of RNA sequences [15]. Notice that the special case where  $|g_1| \leq 1$  and  $|g_2| = 0$  corresponds to usual palindromes, and the special case where  $|g_1| \geq 2$  and  $|g_2| = 0$  corresponds to gapped palindromes.

In this paper, we consider a new class of generalized palindromes where  $|g_1| \leq 1$  and  $|g_2| \geq 1$ , i.e., palindromes with gaps *inside* one of its arms. We call such palindromes as *single-arm-gapped palindromes* (SAGPs). For instance, string  $\text{abb|ca|cb|bc|bba}$  is an SAGP of this kind, taking  $w = \text{abb}$ ,  $g_1 = \varepsilon$  (the empty string),  $g_2 = \text{ca}$ , and  $u = \text{cb}$ .

We are interested in occurrences of SAGPs as substrings of a given string  $T$ . For simplicity, we will concentrate on SAGPs with  $|g_1| = 0$  containing a gap in their left arms. However, slight modification of all the results proposed in this paper can easily be applied to all the other cases. For any occurrence of an SAGP  $wguu^Rw^R$  beginning at position  $b$  in  $T$ , the position  $b + |wgu| - 1$  is called the *pivot* of the occurrence of this SAGP. This paper proposes various algorithms to solve the problem of computing longest SAGPs for every pivot in a given string  $T$  of length  $n$ . We classify longest SAGPs into two groups: those which have  $uu^R$  as a maximal palindrome (*type-1*), and the others (*type-2*). Firstly, we show a naïve  $O(n^2)$ -time algorithm for computing type-1 longest SAGPs. Secondly, we present a simple but practical  $O(n^2)$ -time algorithm for computing type-1 longest SAGPs based on simple scans over the suffix array [14]. We also show that the running time of this algorithm can be improved by using a dynamic predecessor/successor data structure. If we employ the van Emde Boas tree [16], we achieve  $O((n + \text{occ}_1) \log \log n)$ -time solution, where  $\text{occ}_1$  is the number of type-1 longest SAGPs to output. Finally, we present an  $O(n + \text{occ}_1)$ -time solution based on the suffix tree [17]. For type-2 longest SAGPs, we show an  $O(n + \text{occ}_2)$ -time algorithm, where  $\text{occ}_2$  is the number of type-2 longest SAGPs to output. Combining the last two results, we obtain an optimal  $O(n + \text{occ})$ -time algorithm for computing all longest SAGPs, where  $\text{occ}$  is the number of outputs.

We performed preliminary experiments to compare practical performances of our algorithms for finding type-1 longest SAGPs; the naïve algorithm, the  $O(n^2)$ -time suffix array based algorithm, and the improved suffix array based algorithm with several kinds of predecessor/successor data structures.

All proofs omitted due to lack of space can be found in Appendix.

**Related work.** For a *fixed* gap length  $d$ , one can find all gapped palindromes  $wgw^R$  with  $|g| = d$  in the input string  $T$  of length  $n$  in  $O(n)$  time [9]. Kolpakov and Kucherov [12] showed an  $O(n + L)$ -time algorithm to compute *long-armed palindromes* in  $T$ , which are gapped palindromes  $wgw^R$  such that  $|w| \geq |g|$ . Here,  $L$  denotes the number of outputs. They also showed how to compute, in  $O(n + L)$  time, *length-constrained palindromes* which are gapped palindromes  $wgw^R$  such that the gap length  $|g|$  is in a predefined range. Recently, Fujishige et al. [6] proposed online algorithms to compute long-armed palindromes and length-constrained palindromes from a given string  $T$ . A gapped palindrome  $wgw^R$  is an  $\alpha$ -gapped palindrome, if  $|wg| \leq \alpha|w|$  for  $\alpha \geq 1$ . Gawrychowski et al. [7] showed that the maximum number of  $\alpha$ -gapped palindromes occurring in a string of length  $n$  is at most  $28\alpha n + 7n$ . Since long-armed palindromes are 2-gapped palindromes for  $\alpha = 2$ ,  $L = O(n)$  and thus Kolpakov and Kucherov's algorithm runs in  $O(n)$  time. Gawrychowski et al. [7] also proposed an  $O(\alpha n)$ -time algorithm to compute all  $\alpha$ -gapped palindromes in a given string for any predefined  $\alpha \geq 1$ . We emphasize that none of the above algorithms can directly be applied to computing SAGPs.

## 2 Preliminaries

Let  $\Sigma = \{1, \dots, \sigma\}$  be an integer alphabet of size  $\sigma$ . An element of  $\Sigma^*$  is called a *string*. For any string  $w$ ,  $|w|$  denotes the length of  $w$ . The empty string is denoted by  $\varepsilon$ . Let  $\Sigma^+ = \Sigma^* - \{\varepsilon\}$ . For any  $1 \leq i \leq |w|$ ,  $w[i]$  denotes the  $i$ -th symbol of  $w$ . For a string  $w = xyz$ , strings  $x$ ,  $y$ , and  $z$  are called a *prefix*, *substring*, and *suffix* of  $w$ , respectively. The substring of  $w$  that begins at position  $i$  and ends at position  $j$  is denoted by  $w[i..j]$  for  $1 \leq i \leq j \leq |w|$ , i.e.,  $w[i..j] = w[i] \cdots w[j]$ . For  $j > i$ , let  $w[i..j] = \varepsilon$  for convenience. For two strings  $X$  and  $Y$ , let  $\text{lcp}(X, Y)$  denote the length of the longest common prefix of  $X$  and  $Y$ .

For any string  $x$ , let  $x^R$  denote the reversed string of  $x$ , i.e.  $x^R = x[|x|] \cdots x[1]$ . A string  $p$  is called a *palindrome* if  $p = p^R$ . Let  $T$  be any string of length  $n$ . Let  $p = T[b..e]$  be a palindromic substring of  $T$ . The position  $i = \lfloor \frac{b+e}{2} \rfloor$  is called the *center* of this palindromic substring  $p$ . The palindromic substring  $p$  is said to be the *maximal palindrome* centered at  $i$  iff there are no longer palindromes than  $p$  centered at  $i$ , namely,  $T[b-1] \neq T[e+1]$ ,  $b = 1$ , or  $e = n$ .

A string  $x$  is called a *single-arm-gapped palindrome* (SAGP) if  $x$  is in the form of either  $wguu^R w^R$  or  $wucu^R gw^R$ , with some non-empty strings  $w, g, u \in \Sigma^+$  and  $c \in \Sigma \cup \{\varepsilon\}$ . For simplicity and ease of explanations, in what follows we consider only SAGPs whose left arms contain gaps and  $c = \varepsilon$ , namely, those of form  $wguu^R w^R$ . But our algorithms to follow can easily be modified to compute other forms of SAGPs occurring in a string as well.

Let  $b$  be the beginning position of an occurrence of a SAGP  $wguu^R w^R$  in  $T$ , namely  $T[b..b+2|wu|+|g|-1] = wguu^R w^R$ . The position  $i = b + |wgu| - 1$  is called the *pivot* of this occurrence of the SAGP. This position  $i$  is also the center of the palindrome  $uu^R$ . An SAGP  $wguu^R w^R$  for pivot  $i$  in string  $T$  is represented by a quadruple  $(i, |w|, |g|, |u|)$  of integers. In what follows, we will identify the quadruple  $(i, |w|, |g|, |u|)$  with the corresponding SAGP  $wguu^R w^R$  for pivot  $i$ .

For any SAGP  $x = wguu^R w^R$ , let  $\text{armlen}(x)$  denote the length of the arm of  $x$ , namely,  $\text{armlen}(x) = |wu|$ . A substring SAGP  $y = wguu^R w^R$  for pivot  $i$  in a string  $T$  is said to be a *longest SAGP* for pivot  $i$ , if for any SAGP  $y'$  for pivot  $i$  in  $T$ ,  $\text{armlen}(y) \geq \text{armlen}(y')$ .

Notice that there can be different choices of  $u$  and  $w$  for the longest SAGPs at the same pivot. For instance, consider string `ccabcbabbace`. Then,  $(7, 1, 3, 2) = \text{c|abc|ab|ba|c}$  and  $(7, 2, 3, 1) = \text{ca|bca|b|b|ac}$  are both longest SAGPs (with arm length  $|wu| = 3$ ) for the same pivot 7, where the underlines represent the gaps. Of all longest SAGPs for each pivot  $i$ , we regard those that have longest palindromes  $uu^R$  centered at  $i$  as *canonical longest SAGPs* for pivot  $i$ . In the above example,  $(7, 1, 3, 2) = \text{c|abc|ab|ba|c}$  is a canonical longest SAGP for pivot 7, while  $(7, 2, 3, 1) = \text{ca|bca|b|b|ac}$  is not. Let  $\text{SAGP}(T)$  be the set of canonical longest SAGPs for all pivots in  $T$ . In this paper, we present several algorithms to compute  $\text{SAGP}(T)$ .

For an input string  $T$  of length  $n$  over an integer alphabet of size  $\sigma = n^{O(1)}$ , we perform standard preprocessing which replaces all characters in  $T$  with integers from range  $[1, n]$ . Namely, we radix sort the original characters in  $T$ , and replace each original character by its rank in the sorted order. Since the original integer alphabet is of size  $n^{O(1)}$ , the radix sort can be implemented with  $O(1)$  number of bucket sorts, taking

$O(n)$  total time. Thus, whenever we speak of a string  $T$  over an integer alphabet of size  $n^{O(1)}$ , one can regard  $T$  as a string over an integer alphabet of size  $n$ .

**Tools:** Suppose a string  $T$  ends with a unique character that does not appear elsewhere in  $T$ . The *suffix tree* [17] of a string  $T$ , denoted by  $STree(T)$ , is a path-compressed trie which represents all suffixes of  $T$ . Then,  $STree(T)$  can be defined as an edge-labelled rooted tree such that (1) Every internal node is branching; (2) The out-going edges of every internal node begin with mutually distinct characters; (3) Each edge is labelled by a non-empty substring of  $T$ ; (4) For each suffix  $s$  of  $T$ , there is a unique leaf such that the path from the root to the leaf spells out  $s$ . It follows from the above definition of  $STree(T)$  that if  $n = |T|$  then the number of nodes and edges in  $STree(T)$  is  $O(n)$ . By representing every edge label  $X$  by a pair  $(i, j)$  of integers such that  $X = T[i..j]$ ,  $STree(T)$  can be represented with  $O(n)$  space. For a given string  $T$  of length  $n$  over an integer alphabet of size  $\sigma = n^{O(1)}$ ,  $STree(T)$  can be constructed in  $O(n)$  time [5]. For each node  $v$  in  $STree(T)$ , let  $str(v)$  denote the string spelled out from the root to  $v$ . According to Property (4), we sometimes identify each position  $i$  in string  $T$  with the leaf which represents the corresponding suffix  $T[i..n]$ .

Suppose the unique character at the end of string  $T$  is the lexicographically smallest in  $\Sigma$ . The *suffix array* [14] of string  $T$  of length  $n$ , denoted  $SA_T$ , is an array of size  $n$  such that  $SA_T[i] = j$  iff  $T[j..n]$  is the  $i$ th lexicographically smallest suffix of  $T$  for  $1 \leq i \leq n$ . The *reversed suffix array* of  $T$ , denoted  $SA_T^{-1}$ , is an array of size  $n$  such that  $SA_T^{-1}[SA_T[i]] = i$  for  $1 \leq i \leq n$ . The *longest common prefix array* of  $T$ , denoted  $LCP_T$ , is an array of size  $n$  such that  $LCP_T[1] = -1$  and  $LCP_T[i] = lcp(T[SA_T[i-1]..n], T[SA_T[i]..n])$  for  $2 \leq i \leq n$ . The arrays  $SA_T$ ,  $SA_T^{-1}$ , and  $LCP_T$  for a given string  $T$  of length  $n$  over an integer alphabet of size  $\sigma = n^{O(1)}$  can be constructed in  $O(n)$  time [10, 11].

For a rooted tree  $\mathcal{T}$ , the lowest common ancestor  $LCA_{\mathcal{T}}(u, v)$  of two nodes  $u$  and  $v$  in  $\mathcal{T}$  is the deepest node in  $\mathcal{T}$  which has  $u$  and  $v$  as its descendants. It is known that after a linear-time preprocessing on the input tree, querying  $LCA_{\mathcal{T}}(u, v)$  for any two nodes  $u, v$  can be answered in constant time [2].

Consider a rooted tree  $\mathcal{T}$  where each node is either marked or unmarked. For any node  $v$  in  $\mathcal{T}$ , let  $NMA_{\mathcal{T}}(v)$  denote the deepest marked ancestor of  $v$ . There exists a linear-space algorithm which marks any unmarked node and returns  $NMA_{\mathcal{T}}(v)$  for any node  $v$  in amortized  $O(1)$  time [18].

Let  $A$  be an integer array of size  $n$ . A range minimum query  $RMQ_A(i, j)$  of a given pair  $(i, j)$  of indices ( $1 \leq i \leq j \leq n$ ) asks an index  $k$  in range  $[i, j]$  which stores the minimum value in  $A[i..j]$ . After  $O(n)$ -time preprocessing on  $A$ ,  $RMQ_A(i, j)$  can be answered in  $O(1)$  time for any given pair  $(i, j)$  of indices [2].

Let  $S$  be a set of  $m$  integers from universe  $[1, n]$ , where  $n$  fits in a single machine word. A predecessor (resp. successor) query for a given integer  $x$  to  $S$  asks the largest (resp. smallest) value in  $S$  that is smaller (resp. larger) than  $x$ . Let  $u(m, n)$ ,  $q(m, n)$  and  $s(m, n)$  denote the time for updates (insertion/deletion) of elements, the time for predecessor/successor queries, and the space of a dynamic predecessor/successor data structure. Using a standard balanced binary search tree, we have  $u(m, n) = q(m, n) = O(\log m)$  time and  $s(m, n) = O(m)$  space. The Y-fast trie [19] achieves  $u(m, n) = q(m, n) = O(\log \log m)$  expected time and  $s(m, n) = O(m)$  space, while the van Emde

Boas tree [16] does  $u(m, n) = q(m, n) = O(\log \log m)$  worst-case time and  $s(n, m) = O(n)$  space.

### 3 Algorithms for computing canonical longest SAGPs

In this section, we present several algorithms to compute the set  $SAGP(T)$  of canonical longest SAGPs for all pivots in a given string  $T$ .

A position  $i$  in string  $T$  is said to be of *type-1* if there exists a SAGP  $wguu^Rw^R$  such that  $uu^R$  is the maximal palindrome centered at position  $i$ , and is said to be of *type-2* otherwise. For example, consider  $T = \text{baaabaabaacbaabaabac}$  of length 20. Position 13 of  $T$  is of type-1, since there are canonical longest SAGPs  $(13, 4, 4, 2) = \text{abaa|baac|ba|ab|aaba}$  and  $(13, 4, 1, 2) = \text{abaa|c|ba|ab|aaba}$  for pivot 13, where  $\text{ba|ab}$  is the maximal palindrome centered at position 13. On the other hand, Position 6 of  $T$  is of type-2; the maximal palindrome centered at position 6 is  $\text{aaba|abaa}$  but there are no SAGPs in the form of  $wgaaba|abaa^Rw^R$  for pivot 6. The canonical longest SAGPs for pivot 6 is  $(6, 1, 1, 3) = \text{a|a|aba|aba|a}$ .

Let  $Pos_1(T)$  and  $Pos_2(T)$  be the sets of type-1 and type-2 positions in  $T$ , respectively. Let  $SAGP(T, i)$  be the subset of  $SAGP(T)$  whose elements are canonical longest SAGPs for pivot  $i$ . Let  $SAGP_1(T) = \bigcup_{i \in Pos_1(T)} SAGP(T, i)$  and  $SAGP_2(T) = \bigcup_{i \in Pos_2(T)} SAGP(T, i)$ . Clearly  $SAGP_1(T) \cup SAGP_2(T) = SAGP(T)$  and  $SAGP_1(T) \cap SAGP_2(T) = \emptyset$ . The following lemma gives an useful property to characterize the type-1 positions of  $T$ .

**Lemma 1.** *Let  $i$  be any type-1 position of a string  $T$  of length  $n$ . Then, a SAGP  $wguu^Rw^R$  is a canonical longest SAGP for pivot  $i$  iff  $uu^R$  is the maximal palindrome centered at  $i$  and  $w^R$  is the longest non-empty prefix of  $T[i + |u^R| + 1..n]$  such that  $w$  occurs at least once in  $T[1..i - |u| - 1]$ .*

We define two arrays  $Pals$  and  $LMost$  as follows:

$$Pals[i] = \{r \mid T[i - r + 1..i + r] \text{ is a maximal palindrome in } T \text{ for pivot } i\}.$$

$$LMost[c] = \min\{i \mid T[i] = c\} \text{ for } c \in \Sigma.$$

By Lemma 1, a position  $i$  in  $T$  is of type-1 iff  $LMost[i + Pals[i] + 1] < i - Pals[i]$ .

**Lemma 2.** *Given a string  $T$  of length  $n$  over an integer alphabet of size  $n^{O(1)}$ , we can determine whether each position  $i$  of  $T$  is of type-1 or type-2 in a total of  $O(n)$  time and space.*

By Lemma 1 and Lemma 2, we can consider an algorithm to compute  $SAGP(T)$  by computing  $SAGP_1(T)$  and  $SAGP_2(T)$  separately, as shown in Algorithm 1. In this algorithm, we also construct an auxiliary array  $NextPos$  defined by  $NextPos[i] = \min\{j \mid i < j, T[i] = T[j]\}$  for each  $1 \leq i \leq n$ , which will be used in Section 3.2.

**Lemma 3.** *Algorithm 1 correctly computes  $SAGP(T)$ .*

In the following subsections, we present algorithms to compute  $SAGP_1(T)$  and  $SAGP_2(T)$  respectively, assuming that the arrays  $Pals$ ,  $LMost$  and  $NextPos$  have already been computed.

---

**Algorithm 1:** computing  $SAGP(T)$ 

---

**Input:** string  $T$  of length  $n$   
**Output:**  $SAGP(T)$

```
1 compute  $Pals$ ;                                /* Algorithm 3 in Appendix */
2 for  $i = n$  downto 1 do
3    $c = T[i]$ ;  $NextPos[i] = LMost[c]$ ;  $LMost[c] = i$ ;
4 for  $i = 1$  to  $n$  do
5   if  $LMost[i + Pals[i] + 1] < i - Pals[i]$  then
6      $Pos_1(T) = Pos_1(T) \cup \{i\}$ ;          /* position  $i$  is of type-1 */
7   else
8      $Pos_2(T) = Pos_2(T) \cup \{i\}$ ;          /* position  $i$  is of type-2 */
9 compute  $SAGP_1(T)$ ;                          /* Section 3.1 */
10 compute  $SAGP_2(T)$ ;                          /* Section 3.2 */
```

---

### 3.1 Computing $SAGP_1(T)$ for type-1 positions

In what follows, we present several algorithms corresponding to the line 9 in Algorithm 1. Lemma 1 allows us greedy strategies to compute the longest prefix  $w^R$  of  $T[i + Pals[i] + 1..n]$  such that  $w$  occurs in  $T[1..i - Pals[i] - 1]$ .

#### Naïve quadratic-time algorithm with RMQs.

Let  $T' = T\$T^R\#$ . We construct the suffix array  $SA_{T'}$ , the reversed suffix array  $SA_{T'}^{-1}$ , and the LCP array  $LCP_{T'}$  for  $T'$ .

For each  $Pals[i]$  in  $T$ , for each gap size  $G = 1, \dots, i - Pals[i] - 1$ , we compute  $W = lcp(T[1..i - Pals[i] - G]^R, T[i + Pals[i] + 1..n])$  in  $O(1)$  time by an RMQ on the LCP array  $LCP_{T'}$ . Then, the gap sizes  $G$  with largest values of  $W$  give all longest SAGPs for pivot  $i$ . Since we test  $O(n)$  gap sizes for every pivot  $i$ , it takes a total of  $O(n^2)$  time to compute  $SAGP_1(T)$ . The working space of this method is  $O(n)$ .

#### Simple quadratic-time algorithm based on suffix array.

Given a string  $T$ , we construct  $SA_{T'}$ ,  $SA_{T'}^{-1}$ , and  $LCP_{T'}$  for string  $T' = T\$T^R\#$  as in the previous subsection. Further, for each position  $n + 2 \leq j \leq 2n + 1$  in the reversed part  $T^R$  of  $T' = T\$T^R\#$ , let  $op(j)$  denote its “original” position in the string  $T$ , namely, let  $op(j) = 2n - j + 2$ . Let  $e$  be any entry of  $SA_{T'}$  such that  $SA_{T'}[e] \geq n + 2$ . We associate each such entry of  $SA_{T'}$  with  $op(SA_{T'}[e])$ .

Let  $SA_{T'}[k] = i + Pals[i] + 1$ , namely, the  $k$ th entry of  $SA_{T'}$  corresponds to the suffix  $T[i + Pals[i] + 1..n]$  of  $T$ . Now, the task is to find the longest prefix  $w^R$  of  $T[i + Pals[i] + 1..n]$  such that  $w$  occurs completely inside  $T[1..i - Pals[i] - 1]$ . Let  $b = i - Pals[i] + 1$ , namely,  $b$  is the beginning position of the maximal palindrome  $uu^R$  centered at  $i$ . We can find  $w$  for any maximal SAGP  $wguu^Rw^R$  for pivot  $i$  by traversing  $SA_{T'}$  from the  $k$ th entry forward and backward, until we encounter the nearest entries  $p < k$  and  $q > k$  on  $SA_{T'}$  such that  $op(SA_{T'}[p]) < b - 1$  and  $op(SA_{T'}[q]) < b - 1$ , if they



exist. The size  $W$  of  $w$  is equal to

$$\max\{\min\{LCP_{T'}[p+1], \dots, LCP_{T'}[k]\}, \min\{LCP_{T'}[k+1], \dots, LCP_{T'}[q]\}\}. \quad (1)$$

Assume w.l.o.g. that  $p$  gives a larger lcp value with  $k$ , i.e.  $W = \min\{LCP_{T'}[p+1], \dots, LCP_{T'}[k]\}$ . Let  $s$  be the largest entry of  $SA_{T'}$  such that  $s < p$  and  $LCP_{T'}[s+1] < W$ . Then, any entry  $t$  of  $SA_{T'}$  such that  $s < t \leq p+1$  and  $op(SA_{T'}[t]) < b-1$  corresponds to an occurrence of a longest SAGP for pivot  $i$ , with gap size  $b - op(SA_{T'}[t]) - 1$ . We output longest SAGP  $(i, W, b - op(SA_{T'}[t]) - 1, |u|)$  for each such  $t$ . The case where  $q$  gives a larger lcp value with  $k$ , or  $p$  and  $q$  give the same lcp values with  $k$  can be treated similarly.

We find  $p$  and  $s$  by simply traversing  $SA_{T'}$  from  $k$ . Since the distance from  $k$  to  $s$  is  $O(n)$ , the above algorithm takes  $O(n^2)$  time. The working space is  $O(n)$ .

### Algorithm based on suffix array and predecessor/successor queries.

Let  $occ_1 = |SAGP_1(T)|$ . For any position  $r$  in  $T$ , we say that the entry  $j$  of  $SA_{T'}$  is *active* w.r.t.  $r$  iff  $op(SA_{T'}[j]) < r - 1$ . Let  $Active(r)$  denote the set of active entries of  $SA_{T'}$  for position  $r$ , namely,  $Active(r) = \{j \mid op(SA_{T'}[j]) < r - 1\}$ .

Let  $t_1 = p$ , and let  $t_2, \dots, t_h$  be the decreasing sequence of entries of  $SA_{T'}$  which correspond to the occurrences of longest SAGPs for pivot  $i$ . Notice that for all  $1 \leq \ell \leq h$  we have  $op(SA_{T'}[t_\ell]) < b - 1$  and hence  $t_\ell \in Active(b)$ , where  $b = i - |u| + 1$ . Then, finding  $t_1$  reduces to a predecessor query for  $k$  in  $Active(b)$ . Also, finding  $t_\ell$  for  $2 \leq \ell \leq h$  reduces to a predecessor query for  $t_{\ell-1}$  in  $Active(b)$ .

To effectively use the above observation, we compute an array  $U$  of size  $n$  from  $Pals$  such that  $U[b]$  stores a list of all maximal palindromes in  $T$  which begin at position  $b$  if they exist, and  $U[b]$  is nil otherwise.  $U$  can be computed in  $O(n)$  time e.g., by bucket sort. After computing  $U$ , we process  $b = 1, \dots, n$  in increasing order. Assume that when we process a certain value of  $b$ , we have maintained a dynamic predecessor/successor query data structure for  $Active(b)$ . The key is that the same set  $Active(b)$  can be used to compute the longest SAGPs for every element in  $U[b]$ , and hence we can use the same predecessor/successor data structure for all of them. After processing all elements in  $U[b]$ , we insert all elements of  $Active(b-1) \setminus Active(b)$  to the predecessor/successor data structure. Each element to insert can be easily found in constant time.

Since we perform  $O(n + occ_1)$  predecessor/successor queries and  $O(n)$  insertion operations in total, we obtain the following theorem.

**Theorem 1.** *Given a string  $T$  of size  $n$  over an integer alphabet of size  $\sigma = n^{O(1)}$ , we can compute  $SAGP_1(T)$  in  $O(n(u(n, n) + q(n, n)) + occ_1 \cdot q(n, n))$  time with  $O(n + s(n, n))$  space by using the suffix array and a predecessor/successor data structure, where  $occ_1 = |SAGP_1(T)|$ .*

Since every element of  $Active(b)$  for any  $b$  is in range  $[1, 2n + 2]$ , we can employ the van Emde Boas tree [16] as the dynamic predecessor/successor data structure using  $O(n)$  total space. Thus we obtain the following theorem.

**Theorem 2.** *Given a string  $T$  of size  $n$  over an integer alphabet of size  $\sigma = n^{O(1)}$ , we can compute  $SAGP_1(T)$  in  $O((n + occ_1) \log \log n)$  time and  $O(n)$  space by using the suffix array and the van Emde Boas tree, where  $occ_1 = |SAGP_1(T)|$ .*

### Optimal-time algorithm based on suffix tree.

In this subsection, we show that the problem can be solved in *optimal* time and space, using the following three suffix trees regarding the input string  $T$ . Let  $\mathcal{T}_1 = \text{STree}(T\$T^R\#)$  for string  $T\$T^R\#$  of length  $2n + 2$ , and  $\mathcal{T}_2 = \text{STree}(T^R\#)$  of length  $n + 1$ . These suffix trees  $\mathcal{T}_1$  and  $\mathcal{T}_2$  are static, and thus can be constructed offline, in  $O(n)$  time for an integer alphabet. We also maintain a growing suffix tree  $\mathcal{T}_2' = \text{STree}(T^R[k..n]\#)$  for decreasing  $k = n, \dots, 1$ .

**Lemma 4.** *Given  $\mathcal{T}_2 = \text{STree}(T^R\#)$ , we can maintain  $\mathcal{T}_2' = \text{STree}(T^R[k..n]\#)$  for decreasing  $k = n, \dots, 1$  incrementally, in  $O(n)$  total time for an integer alphabet of size  $n^{O(1)}$ .*

**Theorem 3.** *Given a string  $T$  of length  $n$  over an integer alphabet of size  $\sigma = n^{O(1)}$ , we can compute  $\text{SAGP}_1(T)$  in optimal  $O(n + \text{occ}_1)$  time and  $O(n)$  space by using suffix trees, where  $\text{occ}_1 = |\text{SAGP}_1(T)|$ .*

*Proof.* We first compute the array  $U$ . Consider an arbitrary fixed  $b$ , and let  $uu^R$  be a maximal palindrome stored in  $U[b]$  whose center is  $i = b + |u| - 1$ . Assume that we have a growing suffix tree  $\mathcal{T}_2'$  for string  $T^R[n - b + 1..n]\#$  which corresponds to the prefix  $T[1..b]$  of  $T$  of size  $b$ . We use a similar strategy as the suffix array based algorithms. For each position  $2n - b + 2 \leq j \leq 2n + 1$  in string  $T' = T\$T^R\#$ ,  $1 \leq \text{op}(j) \leq b - 2$ . We maintain the NMA data structure over the suffix tree  $\mathcal{T}_1$  for string  $T'$  so that all the ancestors of the leaves whose corresponding suffixes start at positions  $2n - b + 2 \leq j \leq 2n + 1$  are marked, and any other nodes in  $\mathcal{T}_1$  remain unmarked at this step.

As in the suffix-array based algorithms, the task is to find the longest prefix  $w^R$  of  $T[i + |u^R| + 1..n]$  such that  $w$  occurs completely inside  $T[1..b - 2] = T[1..i - |u| - 1]$ . In so doing, we perform an NMA query from the leaf  $i + |u^R| + 1$  of  $\mathcal{T}_1$ , and let  $v$  be the answer to the NMA query. By the way how we have maintained the NMA data structure, it follows that  $\text{str}(v) = w^R$ .

To obtain the occurrences of  $w$  in  $T[1..b - 2]$ , we switch to  $\mathcal{T}_2'$ , and traverse the subtree rooted at  $v$ . Then, for any leaf  $\ell$  in the subtree,  $(i, |\text{str}(v)|, b - \text{op}(\ell), |u|)$  is a canonical longest SAGP for pivot  $i$  (see also Fig. 10 in Appendix).

After processing all the maximal palindromes in  $U[b]$ , we mark all unmarked ancestors of the leaf  $2n - b$  of  $\mathcal{T}_1$  in a bottom-up manner, until we encounter the lowest ancestor that is already marked. This operation is a preprocessing for the maximal palindromes in  $U[b + 1]$ , as we will be interested in the positions between 1 and  $\text{op}(2n - b) = b - 1$  in  $T$ . In this preprocessing, each unmarked node is marked at most once, and each marked node will remain marked. In addition, we update the growing suffix tree  $\mathcal{T}_2'$  by inserting the new leaf for  $T^R[n - b..n]\#$ .

We analyze the time complexity of this algorithm. Since all maximal palindromes in  $U[b]$  begin at position  $b$  in  $T$ , we can use the same set of marked nodes on  $\mathcal{T}_1$  for all of those in  $U[b]$ . Thus, the total cost to update the NMA data structure for all  $b$ 's is linear in the number of unmarked nodes that later become marked, which is  $O(n)$  overall. The cost for traversing the subtree of  $\mathcal{T}_2'$  to find the occurrences of  $w$  can be charged to the number of canonical longest SAGPs to output for each pivot, thus it takes  $O(\text{occ}_1)$  time for all pivots. Updating the growing suffix tree  $\mathcal{T}_2'$  takes overall  $O(n)$  time by Lemma 4.



What remains is how to efficiently link the new internal node introduced in the growing suffix tree  $\mathcal{T}'_2$ , to its corresponding node in the static suffix tree  $\mathcal{T}_1$  for string  $T'$ . This can be done in  $O(1)$  time using a similar technique based on LCA queries on  $\mathcal{T}_1$ , as in the proof of Lemma 4. Summing up all the above costs, we obtain  $O(n + occ_1)$  optimal running time and  $O(n)$  working space.  $\square$

### 3.2 Computing $SAGP_2(T)$ for type-2 positions

In this subsection, we present an algorithm to compute  $SAGP_2(T)$  in a given string  $T$ , corresponding to the line 10 in Algorithm 1.

**Lemma 5.** *Every (not necessarily longest) SAGP for pivot  $i$  must end at one of the positions between  $i + 2$  and  $i + Pals[i]$ .*

**Lemma 6.** *For any type-2 position  $i$  in string  $T$ , if  $wguu^Rw^R$  is a canonical longest SAGP for pivot  $i$ , then  $|w| = 1$ .*

For every type-2 position  $i$  in  $T$ , let  $u = T[i..i + Pals[i]]$ . By Lemma 6, any canonical longest SAGP is of the form  $cguu^Rc$  for  $c \in \Sigma$ . For each  $2 \leq k \leq Pals[i]$ , let  $c_k = u^R[k]$ , and let  $u_k^R$  be the proper prefix of  $u^R$  of length  $k - 1$ . Now, observe that the largest value of  $k$  for which  $LMost[c_k] \leq i - |u_k| - 1$  corresponds to a canonical longest SAGP for pivot  $i$ , namely,  $c_k g_k u_k u_k^R c_k$  is a canonical longest SAGP for pivot  $i$ , where  $g_k = T[LMost[c_k] + 1..i - |u_k|]$ . In order to efficiently find the largest value of such, we consider a function  $findR(t, i)$  defined by

$$findR(t, i) = \min\{r \mid t \leq r < i, T[l] = T[r] \text{ for } 1 \leq \exists l < r\} \cup \{+\infty\}.$$

**Lemma 7.** *For any type-2 position  $i$  in  $T$ , quadruple  $(i, 1, r - LMost[T[r]], i - r)$  represents a canonical longest SAGP for pivot  $i$ , where  $r = findR(i - Pals[i] + 1, i) \neq \infty$ . Moreover, its gap is the longest among all the canonical longest SAGPs for pivot  $i$ .*

By Lemma 7, we can compute a canonical longest SAGP for any type-2 pivot  $i$  in  $O(1)$  time, assuming that the function  $findR(t, i)$  returns a value in  $O(1)$  time. We define an array  $FindR$  of size  $n$  by

$$FindR[t] = \min\{r \mid t \leq r, T[l] = T[r] \text{ for } 1 \leq \exists l < r\} \cup \{+\infty\}, \quad (2)$$

for  $1 \leq t \leq n$ . If the array  $FindR$  has already been computed, then  $findR(t, i)$  can be obtained in  $O(1)$  time by  $findR(t, i) = FindR[t]$  if  $FindR[t] < i$ , and  $+\infty$  otherwise.

Algorithm 2 shows a pseudo-code to compute  $FindR$ . Table 2 in Appendix shows an example.

**Lemma 8.** *Algorithm 2 correctly computes the array  $FindR$  in  $O(n)$  time and space.*

By Lemma 8, we can compute  $SAGP_2(T)$  for type-2 positions as follows.

**Theorem 4.** *Given a string  $T$  of length  $n$  over an integer alphabet of size  $n^{O(1)}$ , we can compute  $SAGP_2(T)$  in  $O(n + occ_2)$  time and  $O(n)$  space, where  $occ_2 = |SAGP_2(T)|$ .*

---

**Algorithm 2:** constructing the array *FindR*

---

**Input:** string  $T$  of length  $n$

**Output:** array *FindR* of size  $n$

```
1 Let  $Occ_1$  and  $Occ_2$  be arrays of size  $\Sigma_T$  initialized by  $+\infty$ ;
2 Let FindR be an arrays of size  $n$ , and let Stack be an empty stack;
3  $min_{in} = +\infty$ ;
4 for  $i = n$  downto 1 do
5    $c = T[i]$ ;  $Occ_2[c] = Occ_1[c]$ ;  $Occ_1[c] = i$ ;
6    $min_{in} = \min\{min_{in}, Occ_2[c]\}$ ;
7   Stack.push( $i$ );
8   while Stack is not empty and  $LMost[T[Stack.top]] \geq i$  do Stack.pop()
9    $min_{out} = Stack.top$  if Stack is not empty else  $+\infty$ ;
    $FindR[i] = \min\{min_{in}, min_{out}\}$ 
```

---

*Proof.* For a given  $T$ , we first compute the array *FindR* by Algorithm 2. By Lemma 7, we can get a canonical longest SAGP  $x_1 = (i, 1, |g_1|, Pals[i])$  if any, in  $O(1)$  time by referring to *LMost* and *FindR*. Note that  $x_1$  is the one whose gap  $|g_1|$  is the longest. Let  $b_1 = i - Pals[i] - |g_1|$  be the beginning position of  $x_1$  in  $T$ . Then the next shorter canonical longest SAGP for the same pivot  $i$  begins at position  $b_2 = NextPos[b_1]$ . By repeating this process  $b_{j+1} = NextPos[b_j]$  while the gap size  $|g_j| = i - Pals[i] - |b_j|$  is positive, we obtain all the canonical longest SAGPs for pivot  $i$ . Overall, we can compute all canonical longest SAGPs for all pivots in  $T$  in  $O(n + occ_2)$  time. The space requirement is clearly  $O(n)$ .  $\square$

We now have the main theorem from Theorem 3, Lemma 2, Lemma 3, and Theorem 4 as follows.

**Theorem 5.** *Given a string  $T$  of length  $n$  over an integer alphabet of size  $n^{O(1)}$ , Algorithm 1 can compute  $SAGP(T)$  in optimal  $O(n + occ)$  time and  $O(n)$  space, where  $occ = |SAGP(T)|$ .*

## 4 Experiments

In this section, we show some experimental results which compare performance of our algorithms for computing  $SAGP_1(T)$ . We implemented the naïve quadratic-time algorithm (Naïve), the simple quadratic-time algorithm which traverses suffix arrays (Traverse), and three versions of the algorithm based on suffix array and predecessor/successor data structure, each employing red-black trees (RB tree), Y-fast tries (Y-fast trie), and van Emde Boas trees<sup>1</sup> (vEB tree), as the predecessor/successor data structure.

We implemented all these algorithms with Visual C++ 12.0 (2013), and performed all experiments on a PC (Intel© Xeon CPU W3565 3.2GHz, 12GB of memory) running

---

<sup>1</sup>We modified a van Emde Boas tree implementation from <https://code.google.com/archive/p/libveb/> so it works with Visual C++.

on Windows 7 Professional. In each problem, we generated a string randomly and got the average time for ten times attempts.

Table 1: Running times (in milli-sec.) on randomly generated strings of length 10000, 50000, and 100000 with  $|\Sigma| = 10$ .

$n$	Naïve	Traverse	RB tree	vEB tree	Y-fast trie
10000	247.2	3.8	6.3	85.7	11.7
50000	7661.0	18.6	37.2	128.9	62.6
100000	32933.2	38.7	80.3	191.9	133.7

We tested all programs on strings of lengths 10000, 50000, and 100000, all from an alphabet of size  $|\Sigma| = 10$ . Table 1 shows the results. From Table 1, we can confirm that **Traverse** is the fastest, while **Naïve** is by far the slowest. We further tested the algorithms on larger strings with  $|\Sigma| = 10$ . In this comparison, we excluded **Naïve** as it is too slow. The results are shown in Fig. 1. As one can see, **Traverse** was the fastest for all lengths. We also conducted the same experiments varying alphabet sizes as 2, 4, and 20, and obtained similar results as the case of alphabet size 10.

To verify why **Traverse** runs fastest, we measured the average numbers of suffix array entries which are traversed, per pivot and output (i.e., canonical longest SAGP). Fig. 2 shows the result. We can observe that although in theory  $O(n)$  entries can be traversed per pivot and output for a string of length  $n$ , in both cases the actual number is far less than  $O(n)$  and grows very slowly as  $n$  increases. This seems to be the main reason why **Traverse** is faster than **RB tree**, **vEB tree**, and **Y-fast trie** which use sophisticated but also complicated predecessor/successor data structures.

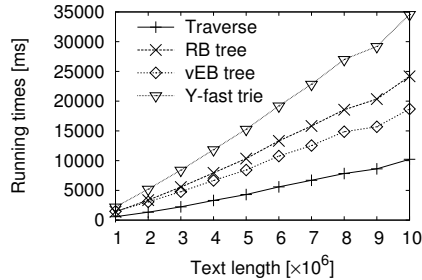


Figure 1: Running times (in milli-sec.) on randomly generated strings of length from 1000000 to 10000000 with  $|\Sigma| = 10$ .

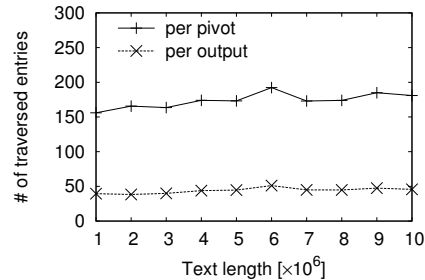


Figure 2: Average numbers of traversed entries of suffix array per pivot and per output on randomly generated strings.

## References

- [1] A. Apostolico, D. Breslauer, and Z. Galil. Parallel detection of all palindromes in a string. *Theor. Comput. Sci.*, 141(1&2):163–173, 1995.
- [2] M. A. Bender and M. Farach-Colton. The LCA problem revisited. In *LATIN*, pages 88–94, 2000.

- [3] M. Crochemore, L. Ilie, C. S. Iliopoulos, M. Kubica, W. Rytter, and T. Walen. Computing the longest previous factor. *Eur. J. Comb.*, 34(1):15–26, 2013.
- [4] X. Droubay and G. Pirillo. Palindromes and Sturmian words. *Theor. Comput. Sci.*, 223(1-2):73–85, 1999.
- [5] M. Farach-Colton, P. Ferragina, and S. Muthukrishnan. On the sorting-complexity of suffix tree construction. *J. ACM*, 47(6):987–1011, 2000.
- [6] Y. Fujishige, M. Nakamura, S. Inenaga, H. Bannai, and M. Takeda. Finding gapped palindromes online. In *IWOCA 2016*, 2016. To appear.
- [7] P. Gawrychowski, T. I, S. Inenaga, D. Köppl, and F. Manea. Efficiently finding all maximal  $\alpha$ -gapped repeats. In *STACS 2016*, pages 39:1–39:14, 2016.
- [8] A. Glen, J. Justin, S. Widmer, and L. Q. Zamboni. Palindromic richness. *Eur. J. Comb.*, 30(2):510–531, 2009.
- [9] D. Gusfield. *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press, 1997.
- [10] J. Kärkkäinen, P. Sanders, and S. Burkhardt. Linear work suffix array construction. *J. ACM*, 53(6):918–936, 2006.
- [11] T. Kasai, G. Lee, H. Arimura, S. Arikawa, and K. Park. Linear-time longest-common-prefix computation in suffix arrays and its applications. In *CPM 2001*, pages 181–192, 2001.
- [12] R. Kolpakov and G. Kucherov. Searching for gapped palindromes. *Theor. Comput. Sci.*, 410(51):5365–5373, 2009.
- [13] G. K. Manacher. A new linear-time on-line algorithm for finding the smallest initial palindrome of a string. *J. ACM*, 22(3):346–351, 1975.
- [14] U. Manber and E. W. Myers. Suffix arrays: A new method for on-line string searches. *SIAM J. Comput.*, 22(5):935–948, 1993.
- [15] Y.-Z. Shi, F.-H. Wang, Y.-Y. Wu, and Z.-J. Tan. A coarse-grained model with implicit salt for RNAs: Predicting 3D structure, stability and salt effect. *The Journal of Chemical Physics*, 141(10), 2014.
- [16] P. van Emde Boas. Preserving order in a forest in less than logarithmic time. In *FOCS*, pages 75–84, 1975.
- [17] P. Weiner. Linear pattern matching algorithms. In *14th Annual Symposium on Switching and Automata Theory*, pages 1–11, 1973.
- [18] J. Westbrook. Fast incremental planarity testing. In *ICALP*, pages 342–353, 1992.
- [19] D. E. Willard. Log-logarithmic worst-case range queries are possible in space  $\Theta(N)$ . *Information Processing Letters*, 17:81–84, 1983.

## Appendix

### A Proofs

Here, we present proofs that are omitted due to lack of space.

#### A.1 Proof of Lemma 1

*Proof.* ( $\Rightarrow$ ) Assume on the contrary that  $uu^R$  is not the maximal palindrome centered at  $i$ , and let  $xuu^Rx^R$  be the maximal palindrome centered at position  $i$  with  $|x| \geq 1$ . If  $w^R = x^R$ , then since position  $i$  is of type-1, there must be a SAGP  $w'g'xuu^Rx^Rw'^R$  with  $|w'| \geq 1$  for pivot  $i$ , but this contradicts that  $wguu^Rw^R$  is a longest SAGP for pivot  $i$ . Hence  $x^R$  is a proper prefix of  $w^R$ . See Fig. 3. Let  $x^Rw''^R = w^R$ . Since  $w''^R$  is a non-empty suffix of  $w^R$ ,  $w''$  is a non-empty prefix of  $w$ . This implies that there exists a SAGP  $w''g''xuu^Rx^Rw''^R$  for pivot  $i$ . However, this contradicts that  $wguu^Rw^R$  is a canonical longest SAGP for pivot  $i$ . Consequently,  $uu^R$  is the maximal palindrome centered at  $i$ , and now it immediately follows that  $w^R$  is the longest non-empty prefix of  $T[i + |u^R| + 1..n]$  such that  $w$  occurs at least once in  $T[1..i - |u| - 1]$ .

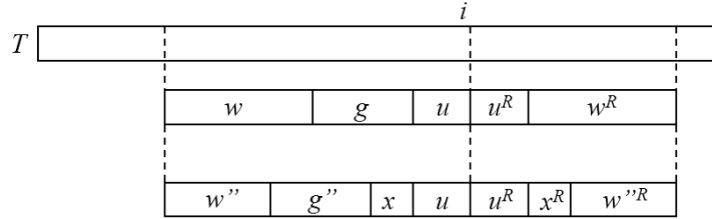


Figure 3: Illustration for a necessary condition for a canonical longest SAGP (proof of ( $\Rightarrow$ ) for Lemma 1):  $wguu^Rw^R$  is a canonical longest SAGP for pivot  $i$ . For the same pivot  $i$ , there cannot exist a SAGP  $w''g''xuu^Rx^Rw''^R$  where  $xuu^Rx^R$  is the maximal palindrome centered at  $i$  and  $w''$  is a prefix of  $w$ , since it contradicts that  $wguu^Rw^R$  is a canonical longest SAGP for  $i$ .

( $\Leftarrow$ ) First, we show that  $wguu^Rw^R$  is a longest SAGP for pivot  $i$ . See Fig. 4. Let  $u'$  be any proper suffix of  $u$ , and assume on the contrary that there exists a SAGP  $w'g'u'u'^Rw'^R$  for pivot  $i$  such that  $|w'u'| > |wu|$ . Since  $|u'| < |u|$ , the occurrence of  $w^R$  at position  $i + |u^R|$  is completely contained in the occurrence of  $w'^R$  at position  $i + |u'^R|$ . This implies that any occurrence of  $w'$  to the left of  $u'u'^R$  completely contains an occurrence of  $w$ , reflected from the occurrence of  $w^R$  in  $w'^R$ . However, the character  $a$  that immediately precedes the occurrence of  $w$  in  $w'$  must be distinct from the character  $b$  that immediately follows  $w^R$ , namely  $a \neq b$ . This contradicts that  $w'g'u'u'^Rw'^R$  is a SAGP for pivot  $i$ . Hence,  $wguu^Rw^R$  is a longest SAGP for pivot  $i$ . Since  $uu^R$  is the maximal palindrome centered at  $i$ , we cannot extend  $u$  to its left nor  $u^R$  to its right for the same center  $i$ . Thus,  $wguu^Rw^R$  is a canonical longest SAGP for pivot  $i$ .  $\square$

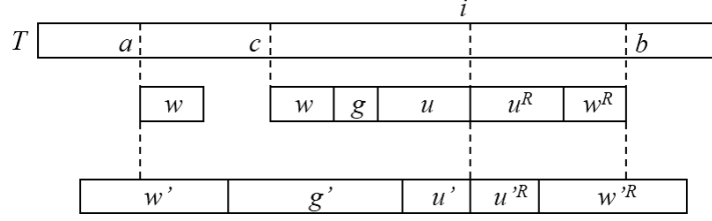


Figure 4: Illustration for a sufficient condition for a canonical longest SAGP (proof of  $(\Leftarrow)$  for Lemma 1):  $uu^R$  is the maximal palindrome centered at  $i$  and  $w^R$  is the longest prefix of  $T[i + |u^R w^R| + 1..n]$  such that  $w$  occurs at least once in  $T[1..i - |u| - 1]$ , and thus  $c \neq b$ . Then, there cannot exist a longer SAGP  $w'g'u'u'^R w'^R$  for the same pivot  $i$ , since  $a \neq b$ .

### A.2 Proof of Lemma 2

*Proof.* let  $uu^R$  be the maximal palindrome centered at  $i$ . Observe that, by Lemma 1,  $i$  is a type-1 position iff (1) the character  $a = T[i + |u^R| + 1]$  which immediately follows  $u^R$  occurs in  $T[1..i - |u| - 1]$ , or (2)  $i + |u^R| = n$  (namely,  $uu^R$  is a suffix of  $T$ ). It is trivial to check Case (2) in  $O(1)$  time, so we will concentrate on Case (1). Recall that  $|u^R| = |u| = \text{Pals}[i]$ .

Let  $\Sigma_T$  be the set of distinct characters occurring in  $T$ . We construct an array  $L\text{Most}$  of size  $|\Sigma_T|$  such that for each  $1 \leq j \leq |\Sigma_T|$ ,  $L\text{Most}[j]$  stores the leftmost occurrence of the lexicographically  $j$ th character in  $T$ . Using the above observation for Case (1) and the array  $L\text{Most}$ , we can determine in  $O(1)$  time whether a given position  $i$  of  $T$  is of type-1 or type-2 by  $L\text{Most}[i + \text{Pals}[i] + 1] < i - \text{Pals}[i]$ . We can sort the characters in  $\Sigma_T$  in  $O(n)$  by constructing  $SA_T$  in  $O(n)$  time and space.  $\square$

### A.3 Proof of Lemma 3

*Proof.* In line 1, we firstly compute an array  $\text{Pals}$ .  $\text{Pals}[i]$  stores radius  $r$  of maximal palindrome centered at  $i$ . We can compute  $\text{Pals}$  in  $O(n)$  time and space applying Manacher's algorithm [13]. We show how to compute  $\text{Pals}$  in Algorithm 3.

In the first **for**-loop, we construct auxiliary arrays  $L\text{Most}$  and  $\text{NextPos}$ . The correctness of the computation of these arrays is obvious. We use  $\text{NextPos}$  when computing  $\text{SAGP}_2$ . In line 7, since we correctly determine which each position of  $T$  is of type-1 or type-2 by Lemma 2, we must compute  $\text{Pos}_1(T)$  and  $\text{Pos}_2(T)$  in the second **for**-loop. Therefore, by referring each element of  $\text{Pos}_1(T)$  and  $\text{Pos}_2(T)$  respectively, we can compute  $\text{SAGP}_1(T)$  and  $\text{SAGP}_2(T)$ , namely  $\text{SAGP}(T)$ .  $\square$

### A.4 Proof of Lemma 4

*Proof.* We also use  $SA_{T^{\#}}$  and  $SA^{-1}$  in our algorithm, and throughout this proof we abbreviate  $SA_{T^{\#}}$  as  $SA$  and  $SA_{T^{\#}}^{-1}$  as  $SA^{-1}$  for simplicity. Let  $PLV$  and  $NLV$  be



---

**Algorithm 3:** computing  $Pals$       */\* proposed by Manacher [13] \*/*

---

**Input:** string  $T$  of length  $n$

**Output:**  $Pals$  */\* $Pals[i]$  stores the maximal even palindrome for pivot  $i$ \*/*

```

1  $Pals[0] = 0;$ 
2  $i = 2; c = 1; r = 0;$ 
3 while  $c \leq n$  do
4    $j = 2 * c - j + 1;$ 
5   while  $T[i] = T[j]$  do
6      $i ++; j --; r ++;$ 
7    $Pals[c] = r;$ 
8    $d = 1;$ 
9   while  $d \leq r$  do
10     $r_l = Pals[c - d];$ 
11    if  $r_l = r - d$  then break  $Pals[c + d] = \min\{r - d, r_l\};$ 
12     $d ++;$ 
13 if  $d > r$  then  $i ++; r = 0$  else  $r = r_l; c = c + d;$ 

```

---

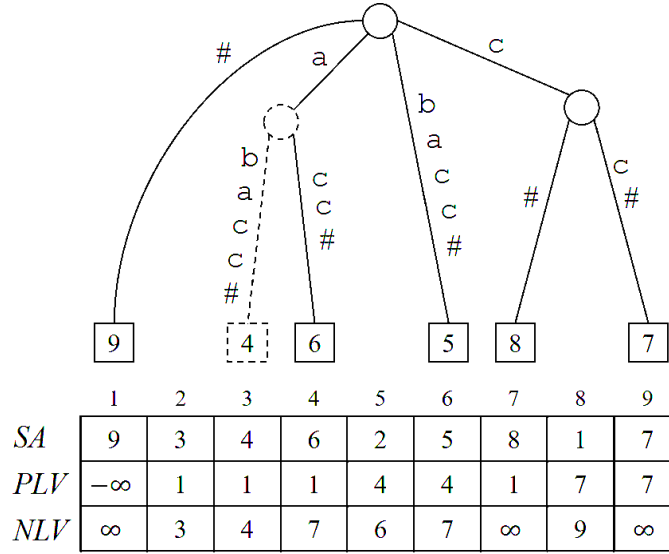


Figure 5: Illustration for the proof of Lemma 4. Consider string  $T = ccababac$ . Here, we illustrate incremental construction of the growing suffix tree  $\mathcal{T}'_2$  for its reversed string  $T^R\# = cbaabacc\#$ . Assume we have constructed  $\mathcal{T}'_2 = STree(T^R[5..8]\#) = STree(bacc\#)$ , and we are to insert a new leaf for the next suffix  $T^R[4..8]\# = abacc\#$  starting at position 4 in  $T^R\#$ .  $SA^{-1}[4] = 3$ , hence we focus on the suffixes of  $T^R\#$  starting at positions  $SA[PLV[3]] = SA[1] = 9$  and  $SA[NLV[3]] = SA[4] = 6$ . Since  $lcp(T^R[9..8]\#, T^R[4..8]\#) = lcp(\#, abacc\#) = 0 < lcp(T^R[4..8]\#, T^R[6..8]\#) = lcp(abacc\#, acc\#) = 1$ , we split the incoming edge of the leaf  $SA[4] = 6$  and insert the new leaf  $SA[3] = 4$  as the left neighbor of the leaf 6.

arrays of size  $n + 1$  each, such that for every  $1 \leq j \leq n + 1$ ,

$$\begin{aligned} PLV[j] &= \max(\{j' \mid 1 \leq j' < j, SA[j'] > SA[j]\} \cup \{-\infty\}), \\ NLV[j] &= \min(\{j' \mid j < j' \leq n + 1, SA[j'] > SA[j]\} \cup \{\infty\}). \end{aligned}$$

Intuitively,  $PLV[j]$  and  $NLV[j]$  indicate the entries of  $SA$  that correspond to the lexicographically closest suffixes to the left and to the right of the suffix  $T^R[SA[j]..n]\#$  which occur positions larger than  $b$ , respectively. If such entries do not exist, then let  $PLV[j] = -\infty$  and  $NLV[j] = \infty$ . See Fig. 5 for concrete examples of  $PLV$  and  $NLV$  arrays.

Suppose we have constructed  $\mathcal{T}'_2 = STree(T^R[k + 1..n]\#)$  up to position  $k + 1$ , and we wish to update it with the new character  $T^R[k]$  at position  $k$ . What is required here is to insert a new leaf corresponding to the suffix  $T^R[k..n]\#$  to the suffix tree. If we use a variant of Weiner's algorithm [17], we can do this in  $O(\log \sigma)$  (amortized) time, but this becomes  $O(\log n)$  for integer alphabets of size  $\sigma = n^{O(1)}$ , and thus it is not enough for our goal. To achieve  $O(1)$  update time per character, we utilize the power of the full suffix tree  $\mathcal{T}_2 = STree(T^R\#)$  and three arrays  $SA$ ,  $SA^{-1}$ ,  $PLV$ , and  $NLV$ .

In what follows, we focus on the case where  $PLV[SA^{-1}[k]] \neq -\infty$  and  $NLV[SA^{-1}[k]] \neq \infty$ . The case where  $PLV[SA^{-1}[k]] = \infty$  or  $NLV[SA^{-1}[k]] = \infty$  is simpler and can be treated similarly. An key observation is that there is a one-to-one correspondence between every leaf of  $STree(T^R[k + 1..n]\#)$  and an entry of  $SA$  which stores a position in  $T^R\#$  which is *larger* than  $k$ . Hence,  $SA[PLV[SA^{-1}[k]]]$  and  $SA[NLV[SA^{-1}[k]]]$  will be, respectively, the left and right neighboring leaves of the new leaf  $k$  in the updated suffix tree  $STree(T^R[k..n]\#)$ .

Given the new position  $k$ , we access  $SA[SA^{-1}[k]]$  which stores the  $k$ th suffix  $T^R[k..n]\#$ . Then we compute the following values  $L$  and  $R$ :

$$\begin{aligned} L &= lcp((T^R\#)[SA[PLV[SA^{-1}[k]]..n + 1]], (T^R\#)[k..n + 1]), \\ R &= lcp((T^R\#)[k..n + 1], (T^R\#)[SA[NLV[SA^{-1}[k]]..n + 1]]). \end{aligned}$$

Depending on the values of  $L$  and  $R$ , we have the following three cases.

- If  $L > R$ , then leaf  $SA[PLV[SA^{-1}[k]]]$  will be the left neighbor of the new leaf  $k$  in the updated suffix tree. Thus, we split the in-coming edge to leaf  $SA[PLV[SA^{-1}[k]]]$  accordingly, and insert the new leaf  $k$ .
- If  $L < R$ , then leaf  $SA[NLV[SA^{-1}[k]]]$  will be the right neighbor of leaf  $k$  in the updated suffix tree. Thus, we split the in-coming edge to leaf  $SA[NLV[SA^{-1}[k]]]$  accordingly, and insert the new leaf  $k$ .
- If  $L = R$ , then leaf  $SA[PLV[SA^{-1}[k]]]$  will be the left neighbor of the new leaf  $k$  and  $SA[NLV[SA^{-1}[k]]]$  will be the right neighbor of leaf  $k$  in the updated suffix tree. Thus, we simply insert the new leaf as a child of the parent of leaves  $SA[PLV[SA^{-1}[k]]]$  and  $SA[NLV[SA^{-1}[k]]]$ .

We then associate the new leaf  $k$  with the  $SA^{-1}[k]$ -th entry of  $SA$  so that later, we can access to this leaf from  $SA[SA^{-1}[k]]$  in  $O(1)$  time. See Fig. 5 for a concrete example on how we insert a new leaf to the growing suffix tree.

Let us analyze the efficiency of our algorithm. Given  $SA$ ,  $PLV[j]$  and  $NLV$  can be constructed in  $O(n)$  time [3]. Then, given a position  $k$  in string  $T^R\#$ , we can access the leaves  $SA[PLV[SA^{-1}[k]]]$  and  $SA[NLV[SA^{-1}[k]]]$  of the full suffix tree  $\mathcal{T}_2 = STree(T^R\#)$  in  $O(1)$  time using  $SA$ ,  $SA^{-1}$ ,  $PLV$ , and  $NLV$  arrays. The values of  $L$  and  $R$  can be computed in  $O(1)$  time by two LCA queries on the full suffix tree  $\mathcal{T}_2$ . In each of the three cases above, it takes  $O(1)$  to insert the new leaf (Notice that we do not have to maintain balanced search trees for branching nodes, since the leaves are sorted by being associated with the corresponding entries of  $SA$ ). Thus, it takes  $O(1)$  time to insert a new leaf to the growing suffix tree  $\mathcal{T}'_2$ . This completes the proof.  $\square$

## A.5 Proof of Lemma 5

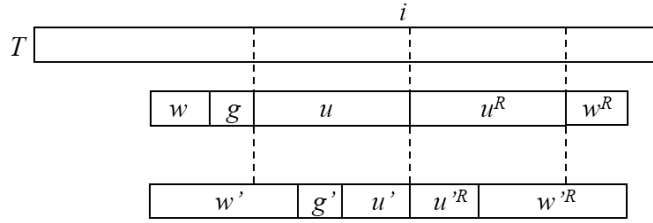


Figure 6: Illustration for Lemma 5.

*Proof.* See Fig. 6. By definition, it is clear that any SAGP for pivot  $i$  must end at position  $i + 2$  or after that. Now, assume on the contrary that there exists a SAGP  $w'g'u'u^Rw'^R$  for pivot  $i$  such that  $i + |u'^Rw'^R| > i + |u^R|$  (it ends after position  $i + |u^R|$ ). Recall that since  $i$  is a type-2 position, we have  $|u'| < |u|$ . Let  $w^R$  be the suffix of  $w'^R$  of size  $|u'^Rw'^R| - |u^R|$ . Then, there exists a SAGP  $wguu^Rw^R$  for pivot  $i$  where  $|g| = |g'|$  and  $uu^R$  is the maximal palindrome centered at  $i$ . However, this contradicts that  $i$  is a type-2 position. Hence, any SAGP for pivot  $i$  must end at position  $i + |u^R|$  or before that.  $\square$

## A.6 Proof of Lemma 6

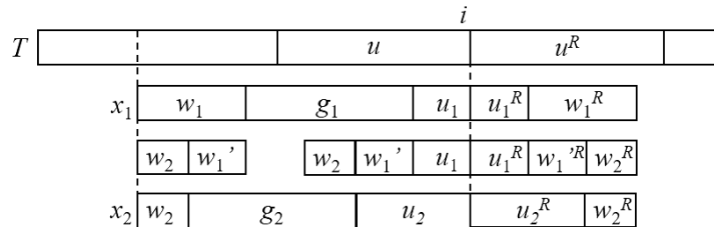


Figure 7: Illustration for Lemma 6.

*Proof.* Let  $x_1 = w_1 g_1 u_1 u_1^R w_1^R$  be a canonical longest SAGP for pivot  $i$ , and on the contrary, suppose that  $|w_1| \geq 2$ . See also Fig 7. Then we can rewrite  $w_1 = w_2 w'_1$  for two non-empty strings  $w_2$  and  $w'_1$ . Let  $uu^R$  be the maximal palindrome centered at  $i$ . Since the position  $i$  is type-2,  $u_1^R w_1^R$  is a prefix of  $u^R$  by Lemma 5, so that  $w_1 u_1$  is a suffix of  $u$ . Moreover, let  $u_2 = w'_1 u_1$  and  $g_2$  be a string satisfying  $g_2 w'_1 = w'_1 g_1$ . Then  $x_2 = w_2 g_2 u_2 u_2^R w_2^R = w_2 g_2 w'_1 u_1 u_2^R w_2^R = w_2 g_2 w'_1 u_1 u_1^R w_1^R w_2^R = w_2 w'_1 g_1 u_1 u_1^R w_1^R w_2^R = w_1 g_1 u_1 u_1^R w_1^R w_2^R = w_1 g_1 u_1 u_1^R w_1^R = x_1$ , that shows  $x_2$  is also a SAGP for pivot  $i$ . Because  $\text{armlen}(x_2) = |w_2 u_2| = |w_1 u_1| = \text{armlen}(x_1)$ ,  $x_2$  is also a *longest* SAGP for pivot  $i$ . Because  $u_2 = w'_1 u_1$  and  $w'_1 \neq \varepsilon$ , we have  $|u_2| < |u_1|$ , which contradicts that  $x_1$  is a *canonical* longest SAGP for pivot  $i$ .  $\square$

## A.7 Proof of Lemma 7

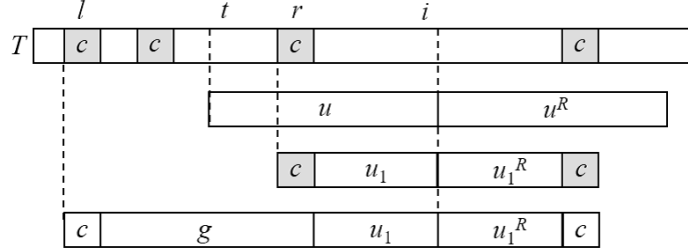


Figure 8: Illustration for Lemma 7.

*Proof.* See Fig. 8. Let  $t$  be the beginning position of  $u$  in  $T$ , namely,  $t = i - |u| + 1$ . Let  $r = \text{findR}(t, i) < \infty$ , and let  $c = T[r]$ . Then by definition of  $\text{findR}(t, i)$ , there exists  $1 \leq l < r$  satisfying  $T[l] = c$ . Therefore,  $x = (i, 1, r - l, i - r)$  is a SAGP for pivot  $i$ . Moreover,  $x$  is *canonical longest* SAGP because  $r$  is minimized, so that  $|u_1| = i - r$  is maximized while  $|w|$  is always 1. Recall that  $L\text{Most}[c]$  is the leftmost position  $l$  satisfying  $T[l] = c$ . Hence, the gap size of the canonical longest SAGP  $(i, 1, r - L\text{Most}[T[r]], i - r)$  is the longest.  $\square$

## A.8 Proof of Lemma 8

*Proof.* The correctness of the computation of  $L\text{Most}$  and  $\text{NextPos}$  is obvious. Let  $\min_{in}[t]$  (resp.  $\min_{out}[t]$ ) be the value of  $\min_{in}$  (reps.  $\min_{out}$ ) when  $i = t$  in the second **for**-loop. We will verify the following loop invariants

$$\min_{out}[t] = \min\{r \mid t \leq r, \ T[l] = T[r] \text{ for } 1 \leq \exists l < t\} \cup \{+\infty\}, \quad (3)$$

$$\min_{in}[t] = \min\{r \mid t \leq r, \ T[l] = T[r] \text{ for } t \leq \exists l < r\} \cup \{+\infty\}, \quad (4)$$

which immediately imply Eq. (2), because  $\text{FindR}[t] = \min\{\min_{out}[t], \min_{in}[t]\}$  in the last line. Eq. (4) is derived from the loop invariants

$$\begin{aligned} \text{Occ}_1[c] &= \min\{j \mid T[j] = c, \ t \leq j\} \cup \{+\infty\}, \\ \text{Occ}_2[c] &= \min\{j \mid T[j] = c, \ \text{Occ}_1[c] < j\} \cup \{+\infty\}, \end{aligned}$$

for any  $c \in \Sigma_T$ . On the other hand, Eq. (3) can be rewritten as

$$\begin{aligned} \min_{out}[t] &= \min\{r \mid t \leq r, \ c = T[r] \text{ for some } c \in \Sigma_T \text{ with } LMost[c] < t\} \cup \{+\infty\} \\ &= \min\{r \mid t \leq r, LMost[T[r]] < t\} \cup \{+\infty\} \\ &= \min\{r \mid LMost[T[r]] < t \leq r\} \cup \{+\infty\}, \end{aligned}$$

and the *Stack* always keeps the values  $\{r \mid LMost[T[r]] < t \leq r\}$  in increasing order from top to bottom. Thus, *Stack.top* returns the minimum value among them. The total running time is  $O(n)$ , because in the *Stack*, each element  $i$  is pushed at most once.  $\square$

## B Examples

Here we present several examples for how our algorithms compute  $SAGP_1(T)$  for a given string.

Consider string  $T = \text{acacabaabca}$  and  $T' = \text{acacabaabca\$acbaabacaca\#}$ , namely,  $T = T\$T^R\#$ . First, we compute  $Pals$  and the array  $U$ . Assume we are now processing position  $b = 6$  in  $T$ , then  $U[6] = \{(6, 9)\}$ , where  $(6, 9)$  represents the maximal palindrome  $T[6..9] = \text{baab}$ . Thus we consider pivot  $i = b + \lceil (9 - 6 + 1)/2 \rceil - 1 = 7$ . We have determined that the position 7 is of type-1 position in constant time, using Lemma 2.

### B.1 Example for suffix array based algorithms

First, we show an example for the algorithm based on the suffix array, and its improved version with predecessor/successor queries.

We construct the suffix array  $SA_{T'}$ , the reversed suffix array  $SA_{T'}^{-1}$ , and the LCP array  $LCP_{T'}$  for  $T'$ . In Fig 9, we show these arrays.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
$SA_{T'}$	12	24	11	23	16	7	5	17	8	21	3	19	1	13	15	6	18	9	10	22	4	20	2	14
$LCP_{T'}$	-1	0	0	1	1	3	1	3	2	1	3	3	5	2	0	4	2	1	0	2	2	2	4	1
$SA_{T'}^{-1}$	13	23	11	21	7	16	6	9	18	19	3	1	14	24	15	5	8	17	12	22	10	20	4	2

Figure 9:  $SA_{T'}$ ,  $LCP_{T'}$  and  $SA_{T'}^{-1}$  for string  $T' = \text{acacabaabca\$acbaabacaca\#}$ .

Let  $k$  be the integer such that  $SA_{T'}[k] = i + \lceil (9 - 6 + 1)/2 \rceil + 1 = 7 + 3 = 10$ , namely  $k = 19$ . This can be obtained from  $SA^{-1}[10] = 19$  (See also Fig 9). To compute the longest  $w$ , we traverse  $SA_{T'}[19]$  forward and backward, until we encounter the nearest entries  $p < k$  and  $q > k$  on  $SA_{T'}$  such that  $op(SA_{T'}[p]) < 5$  and  $op(SA_{T'}[q]) < 5$ . Note that these are equivalent to predecessor/successor queries for 19, respectively. Then, we can find  $p = 10$  and  $q = 20$ . Then, the size  $W$  of  $w$  is computed by

$$W = \max\{\min\{LCP_{T'}[11], \dots, LCP_{T'}[19]\}, \min\{LCP_{T'}[20], \dots, LCP_{T'}[20]\}\},$$

and we obtain  $W = 2$ . In this case,  $q = 20$  gives a larger lcp value with  $k = 19$ . Thus, we output a canonical longest SAGPs  $(7, 2, 3, 2) = \text{ac|}\underline{\text{aca}}|\text{ba|ab|ca}$ . We further traverse  $SA_{T'}$  from the 20th entry backward as long as successive entries  $s$  fulfill  $LCP_{T'}[s+1] \geq W$ . Then, we find  $s = 22$ , thus we output a canonical longest SAGPs  $(7, 2, 1, 2) = \text{ac|}\underline{\text{a}}|\text{ba|ab|ca}$ . We further traverse  $SA_{T'}$  from the 17th entry backward, finally we reach the 24th entry of  $SA_{T'}$ , which is the last entry of the suffix array. Therefore, we finish the process for position  $b = 7$ .

### B.2 Example for suffix tree based algorithm

Next, we show an example for the linear-time algorithm based on the suffix tree.



We first construct the suffix tree  $\mathcal{T}_1 = \text{STree}(T\$T^R\#)$ . Suppose that we have constructed  $\mathcal{T}_2' = \text{STree}(T^R[8..11]\#)$  and marked all ancestors of every leaf  $v$  such that  $19 < v \leq 24$  in  $\mathcal{T}_1$ . In Fig. 10, we show interesting parts of  $\mathcal{T}_1$  and  $\mathcal{T}_2'$ .

To compute the longest  $w$ , we perform an NMA query from the leaf  $i + |u^R| + 1 = 10$  of  $\mathcal{T}_1$ . As can be seen in Fig. 10, we obtain the nearest marked node  $v = \text{NMA}_{\mathcal{T}_1}(10)$ . Thus, we know that  $w^R = \text{ca}$ . Next, we switch from the node  $v$  of  $\mathcal{T}_1$  to its corresponding node  $v'$  of  $\mathcal{T}_2'$  using a link between them. Then, we traverse the subtree rooted at  $v'$  and obtain all occurrences of  $w^R$ , namely  $w^R = T^R[10..11] = T^R[8..9] = \text{ca}$  at positions 10 and 8 in the reversed string  $T^R\#$ . Since  $\text{op}(10) = 2$  and  $\text{op}(8) = 4$ , we obtain the canonical longest SAGPs  $(7, 2, 3, 2) = \text{ac}|\underline{\text{aca}}|\text{ba}|\text{ab}|\text{ca}$ . and  $(7, 2, 1, 2) = \text{ac}|\underline{\text{a}}|\text{ba}|\text{ab}|\text{ca}$  for pivot 7.

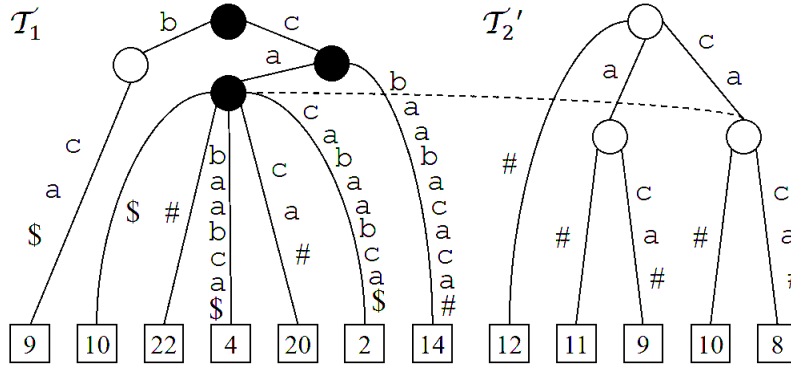


Figure 10: Showing interesting parts of  $\mathcal{T}_1 = \text{STree}(T')$  and  $\mathcal{T}_2' = \text{STree}(T^R[8..11]\#)$ , where  $T = \text{acacabaabca}$ ,  $T^R = \text{acbaabacaca}$  and  $T' = \text{acacabaabca}\$ \text{acbaabacaca}\#$ . In  $\mathcal{T}_1$ , we represent the marked internal nodes by black circles, the unmarked internal nodes by white circles, and the leaves by squares in which the numbers denote the beginning positions of the corresponding suffixes in the string. The dotted line represents the link between the node for string  $c$  in  $\mathcal{T}_1$  and that in  $\mathcal{T}_2'$ .

Table 2: Arrays  $L\text{Most}$ ,  $\text{NextPos}$ , and  $\text{FindR}$  for a string  $T = \text{dbbaacbcbad}$ . For the sake of understanding, we also provide the values of  $\text{min}_{\text{out}}$  and  $\text{min}_{\text{in}}$  in the  $i$ -th loop of Algorithm 2. These values are computed from right to left.

	$L\text{Most}$		1	2	3	4	5	6	7	8	9	10	11
a	4	$T$	d	b	b	a	a	c	b	c	b	a	d
b	2	$\text{NextPos}$	11	3	7	5	10	8	9	$\infty$	$\infty$	$\infty$	$\infty$
c	6	$\text{min}_{\text{out}}$	$\infty$	11	3	7	5	7	7	8	9	10	11
d	1	$\text{min}_{\text{in}}$	3	3	5	5	8	8	9	$\infty$	$\infty$	$\infty$	$\infty$
		$\text{FindR}$	3	3	3	5	5	7	7	8	9	10	11